Multimedia Modeling Using MPEG-7 for Authoring Multimedia Integration

Tien Tran-Thuong, Cécile Roisin

Multimedia Modeling Using MPEG-7 for Authoring Multimedia Integration

Tien Tran-Thuong
WAM Project, INRIA Rhône-Alpes, 655 avenue de l'Europe - Montbonnot - 38334 Saint Ismier - France
Httv, 31, Avenue de Granier - 38240 Meylan - France
(+33) (0)4 76 18 52 74
tien.tran_thuong@inrialpes.fr

Cécile Roisin
WAM Project, INRIA Rhône-Alpes, 655 avenue de l'Europe - Montbonnot - 38334 Saint Ismier - France
Université Pierre Mendès - 38000 Grenoble - France
(+33) (0)4 76 61 53 60
cecile.roisin@inrialpes.fr

ABSTRACT
In this paper, we describe an approach to audiovisual data modeling for multimedia integration and synchronization. The approach chosen consists in using description tools from Multimedia Description Schemes of standard MPEG-7 to describe audiovisual contents and in integrating these description models into a multimedia integration and synchronization model close to SMIL. The resulting model provides relevant specification tools for the fine integration of multimedia fragments into multimedia presentations. An authoring environment illustrates the authoring features that can be obtained thanks to these integrated models.

Categories and Subject Descriptors
D.3.3 [**Information Interfaces And Presentation**]: Multimedia Information Systems – *Evaluation/methodology, Hypertext navigation and maps, Video.*

General Terms
Design, Documentation, Experimentation, Languages.

Keywords
Multimedia document authoring, Fine-grained synchronization, Content description, MPEG-7, Multimedia description schemes.

1. INTRODUCTION
The emerging standards of SVG, SMIL and MPEG-4 provide a new process for authoring and presenting multimedia documents, also known as multimedia integration and synchronization. It introduces a new multimedia type which enables the integration of a rich set of media into more complex structures and provides news interaction capacity in multimedia presentations.

These enhanced features cannot be created with most of the current media production tools like *Adobe Premiere*, *CineKit*, *Vane*, and *Movi2d*. Therefore, the new emerging media production tools such as *Grins*, *Limsee2.0*, *IBM tool kits for MPEG-4*, *X-SMIL*, and *Macromedia Director* can be used for this purpose. These tools provide a sophisticated high level graphical editing interface like *timeline* and *layout* views for integrating and synchronizing a set of media. However all of them still require the author a long and relatively complex authoring process, especially when fine-grained synchronization is desired. As an example: an author wants to display a text introducing a character in a video when this character occurs on screen. The authoring process requires the manual determination of the temporal information, the *begin* time and the *stop* time, of the appearance of the character in the video and then the absolute temporal placement of the text along with this temporal information. The difficulty involves the effort taken to determine temporal information inside the video, because current multimedia authoring tools do not support media content analysis and visualization of high level content structures of video. It becomes more complex when making a hyperlink on the video character or making a text following the video character because the author needs to determine not only temporal information but also spatio-temporal information of the character. At this point, the most important standard multimedia integration model SMIL fails to integrate these specification needs.

As an example of the resulting editing process of this situation, [6] proposes a courseware production model using SMIL in which the authoring of courseware requires to manually cut the video course material into video clips corresponding to the slideshows. Such a costly and tedious work is the consequence of the lack of fine-grained media structure (media content modeling) in SMIL.

The objective of this paper is to propose a way to make the authoring of complex and sophisticated multimedia presentations easier. It provides a solution based on MPEG-7 tools for content modeling and shows how to integrate these tools in a SMIL-like multimedia model for obtaining a complete multimedia authoring tool.

The paper is organized as follows: The basic requirements for multimedia authoring and the need of multimedia content modeling for sophisticated multimedia authoring are discussed in section 2. Section 3 shows that a gap exists between the multimedia content modeling and multimedia integration and synchronization that prevents the direct use of media description data in multimedia authoring. Related works are then discussed in section 4. Section 5 explains our media content description model for multimedia authoring; the expressions of this model using MPEG-7 description schemes are also presented. Sections 6 and 7 present the resulting global multimedia integration model and the

MIR'03, November 7, 2003, Berkeley, California, USA.

experimentations of this work inside the *Mdefi* authoring tool. Finally, section 8 gives some conclusion and perspectives.

2. MULTIMEDIA AUTHORING REQUIREMENTS

Multimedia integration authoring is the composition of a new multimedia presentation from a set of media. The composition is specified in terms of rules issued from a multimedia integration model that allows the expression of different facets of multimedia presentation such as: *Media*, *Temporal*, *Spatial* and *Link*.

- *Media* allows to specify the location of media content;
- *Spatial* allows to specify the layout of each media on screen;
- *Temporal* allows to specify the temporal presentation of each media;
- *Link* allows to set hyperlinks on the media, which creates the navigation and interaction scenario of the presentation.

Figure 1 shows a simple visual way to understand the multimedia integration authoring. Each color bar represents a temporal presentation of a media and each rectangle represents a region on screen on which a media will appear. The temporal integration and synchronization authoring can be considered as the task of placing of the bars on the flow of time. The spatial integration and synchronization authoring can be considered as the spatial layout of the rectangles in the space.

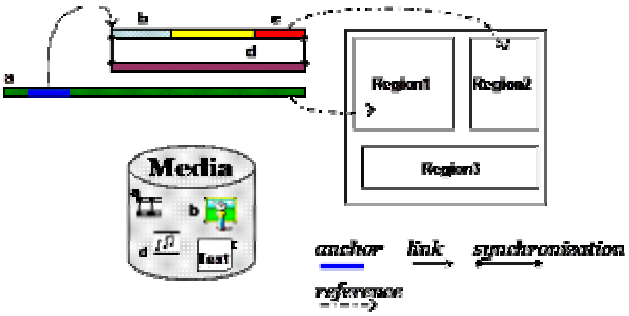

Figure 1. General multimedia integration authoring.

This representation is used as the basic visual authoring approach of the current multimedia authoring tools. However these current authoring tools still cannot support media fragment integration and synchronization authoring features. For instance, how can a temporal interval of an image be placed and resized to synchronize its presentation with a video scene? Or how can an occurrence of a video object be located to make a hypermedia on this moving region? ... And finally what is the cause of these limitations?

The current models are based on a rich set of basic models such as the hierarchical, interval and region-based model [13], known as the most expressive models for expressing the spatio-temporal structure and synchronization in multimedia composition.

An important limitation of the current models is mainly due to the coarse-grained description of the media elements, the smallest granularity of hierarchical structure is media elements. Media elements are thus the black boxes with which compositions are performed (i.e., only the synchronization between the beginning and the end instants of the media are allowed). The problem cannot be solved by simply using the *Anchor* or the *Area* techniques of Hytime, HTML and SMIL models. Indeed they just allow to specify low level media fragments.

```
<video id="#a" src="video/wedingVideo.mpg">
    <area id="lover"
          begin= "5s" end= "7s"          <!—temporal identification of the segment-->
          coords= "84, 249, 188, 272" <!—spatial identification of the segment-->
          href ="#b" />                  <!—hyperlink-->
</video>
```

Figure 2. Specification of a video fragment by using the area element.

Our aim is to go beyond these limits resulting from the use of absolute and non-significant specifications. Usually the media portion that the author wants to synchronize has its own semantic, temporal, and spatial information that can be exploited for its composition. As an example, a moving object in a video has properties such as color, motion, and shape that cannot be expressed by both simple *anchor* and *area* elements, instead it can be perfectly described by multimedia content description tools. Therefore the following question arises: can a multimedia content description be used for multimedia integration authoring? And how the content description can be used for authoring? The rest of the paper will discuss about these subjects.

3. THE GAP BETWEEN CONTENT DESCRIPTION AND MULTIMEDIA AUTHORING

Multimedia content description is an approach to index electronic resources. It provides automatic management of electronic resources in order to make easier and more flexible the use of these electronic resources. At present, media content description tools are mainly used for the information retrieval domain. As stated in the above section, we think that media content description tools can be of high interest in the multimedia authoring field, providing deeply access into the media structure for fine-grained composition.

The main issue preventing media content description tools from being used in the multimedia authoring field is the lack of description standards dedicating to multimedia authoring. A relevant media content description for multimedia composition has to describe temporal, spatial, and spatio-temporal structures of media content, instead of only focusing on metadata or feature description. The recent Multimedia Description Schema (MDS) of the MPEG-7 standard [1] provides the segment description schemes that include useful definitions for our needs Segment DS, StillRegion DS, VideoSegment DS, AudioSegment DS, AuioVisualSegment DS, MultimediaSegment DS, etc. However, the main objective of MPEG-7 is also oriented toward archival and retrieval applications, therefore these tools have to be refined for becoming more relevant to multimedia composition (see section 5).

The issues arise not only from the multimedia data description model as shown above but also from the multimedia integration model. Indeed until now, there is no multimedia integration model that can give support for using the media content descriptions in the authoring of multimedia presentations.

The construction of a fundamental architecture for using media content description in multimedia integration authoring is an important research issue.

4. RELATED WORK

There is now a great deal of work related to multimedia content description and multimedia integration. This section will first discuss this work and then position our approach with regard to it, as well as describing the limitations of other approaches to multimedia fragment integration.

4.1 Multimedia Content Description

A number existing works have considered the application of DC (Dublin Core), RDF (W3C - Resource Description Framework) and MPEG-7 (Multimedia Content Description Interface) [1] to multimedia content description.

The Dublin Core metadata standard is a simple effective element set for generating metadata for describing a wide range of information resources, the semantics of which have been established through consensus by an international, cross-disciplinary group. In [7], a schema for Dublin Core-based video metadata representation has been introduced by J. Hunter & L. Armstrong. More complex applications of Dublin Core are harmonization of Dublin Core with other metadata models such as mixed using of Dublin Core with other metadata vocabularies in RDF metadata or harmonization of MPEG-7 with Dublin Core for multimedia content description.

RDF is a framework for metadata. Its broad goal is to provide a general mechanism being suitable for describing information about any domain. Emphasizing facilities to enable automated processing of Web resources, RDF can be used in a variety of application areas from resource discovery, library catalogs and world-wide directories to syndication and aggregation of news, software, and content to personal collections of music, photos. RDF also provides a simple data model, which can accommodate rich semantic descriptions of multimedia content. For instance, J. Saarela, in [11], introduced a video content model based on RDF. Similarly, J. Hunter & L. Armstrong, in [7], proposed a MPEG-7 description definition language (DDL) based on RDF.

MPEG-7 is a standard for audiovisual information description developed by MPEG (Moving Picture Experts Group) working group. It proposes a set of standard descriptors (Ds) and description schemes (DSs) for describing the audiovisual information and a definition description language (DDL) being a language for defining the new tools in each particular application. Thanks to MPEG-7, computational systems can process further audio and visual information. In fact, in [10], L. Rutledge and P. Schmitz proved the need of a media in format MPEG-7 to improve media fragment integration in Web document. Note that the media fragment integration in Web document can be done until now only with textual document as HTML or XML document[1]. *TV Anytime* with their vision of future digital TV that offers the opportunity to provide value-added interactive services, has also claimed that MPEG-7 collection of descriptors and description schemes is able to fulfill the metadata requirements for *TV Anytime*. Many other projects have chosen MPEG-7 to realize systems that allow users to search, browse, and retrieve audiovisual information much more efficiently than they can do today with text-based search engines.

As more and more audiovisual information becomes available from many sources around the world and people would like to use it for various purposes, there are many standards including but not limited to *RDF Site Summary (RSS)*, *SMPTE Metadata Dictionary*, *EBU P/Meta*, *TV Anytime*. These standards provide basic features or basic tools for developing metadata applications. For specific applications we have to refine these standards. Indeed, it is difficult to have a unique model that can satisfy all requirements of various fields. Thus, hybrid or incorporated approaches are often used in sophisticated applications in [7] where a hybrid approach for an MPEG-7 Description Definition Language (DDL) is proposed. Pushed by the recent XML technology, these proper models can be easily encoded in a flexible way that can interoperate with the other models. For example, CARNet Media on Demand [14] decided to use their proprietary vocabulary for metadata descriptions for building description structures for media and folders. The model of InfoPyramid [9] allows to handle multimedia content description in a multi-abstraction, multi-modal content representation; J. Carrive et al. [2] have proposed a terminological constraint network model based on description logics for managing collections of TV programs.

Our multimedia content description presented here is aimed at a more sophisticated multimedia integration authoring. It is based on MPEG-7 multimedia description schemes (MDS). More about the refinements can be found in section 5.

4.2 Media Fragment Integration and Synchronization

Multimedia integration and synchronization has been largely investigated and well described in the research literature (*MHEG*, *HyTime*, *SMIL*, *ZYX*, *CMIF*, *Firefly*, *Madeus*, etc.). However, few of them have supplied with media fine-grained synchronization. In fact, the *anchor* and *area* technologies are used in these existing standards to decompose media objects into spatial/temporal fragments for hypermedia and fine-grained synchronization. By these ways multimedia fragment integration remains limited to the use of absolute specification and static region, without taking into account the structure of media, and more importantly with few semantic associated with fragments. The last release of the SMIL 2.0 standard provides the use of the *fragment* attribute of the *area* element that can give a more meaningful description of media fragments. However this attribute can only be used for existing structured media such as, HTML, SVG, SMIL or other XML documents. A more important advanced work presented in [10] provides many ways to refer to the MPEG-7 descriptions for media fragment integration. For instance, <video src= "TienWedding.mpg#mpeg7(clip="scene3")"/>, expresses the integration of the MPEG-7 video fragment "scene3". However, the way to author such media fragment integration is yet unsolved, because the identification of a description in a MPEG-7 description resource is not an easy task for the author. A MPEG-7 description resource normally contains huge and rich data from

[1] The SMIL 2.0 model provides the *fragment* attribute that allows referring to any portion of XML document.

which often only a small part is useful for integration. In the above example, only the *begin* and *end* data of the clip video description is interesting to play the clip.

Other works propose more sophisticated composition, but they remain for specific applications. The work in [3] provides a simple way to fine-grained synchronize fragments of continuous media (audio and video) with static documents (text, image). The level of media structure is not deep and rich. It only defines *story*, *clip*, and *scene* elements and the *scene* is the smallest unit of the media structure. It can only provide user interactions from the static media and therefore cannot provide hyperlinks on video fragments.

There is an important limitation in the video-centered hypermedia model, i.e., the logic of the presentation of the document is limited to the sequential logical structure of the continuous narrative media (video or audio). A more generic model for fine-grained composition of multimedia presentation is expected.

The work in [8] has tried to address these problems by the integration of all the analyzed, indexed and composition information inside a multimedia authoring environment. It allows the user to easily create sophisticated Interactive Electronic Technical Manuals. The system uses the *anchor* element of HyTime model to encode extracted fragments. This is a low level way to specify media portions that restricts the reuse of existing media content description.

We propose in this paper a solution for a more meaningful integration of media fragment in multimedia presentations. It is based on a structured media concept as described below.

5. MULTIMEDIA CONTENT MODELING

Among the standards presented in the previous section, MPEG-7 has emerged as the major standard for multimedia content description. Since our aim is to widely use existing multimedia content descriptions, we have chosen to build our description model on this standard. As many standard models, MPEG-7 provides only a set of basic and generic description and descriptor schemas. In order to use correctly this standard, we have firstly identified the features required by our description model for multimedia authoring, and then we have expressed our description model by restricting the MPEG-7 description schemes. In this paper, we present the first step of this experimental model that only covers the structural aspects of media content. In order to illustrate this modeling approach based on a restriction of the MPEG-7 description schemes, we describe the video segment and the moving region models.

5.1 General Model

The media content description model adapted to the composition of multimedia document has to satisfy the following criteria:

1. The multimedia integration authoring is mainly based on the spatial and temporal integration and synchronization of multimedia. The model thus has to focus largely on spatial and temporal structures of the content descriptions.
2. The model has to describe the logical structure (narrative structure) of the media content. The access into the content will be much easier for the author thanks to the narrative structure of the content.
3. Generic synchronizations, for instance, to launch music when a ***red*** object appears, are also necessary to define synchronizations with online media stream as a flow of video survey. Thus the model has to describe the characteristics (e.g., color, texture, shape) and to define the models of these characteristics (e.g., red objects with the shape of a car and moving along a specific path).
4. The conceptual synchronizations are often necessary in composition of multimedia presentation, for example to show the text "My love" over a person each time this one appears on the video. Thus the model has to classify semantically the descriptions such as a set of descriptions of the *lover* on the video.

The above features will allow the integration of multimedia fragments at different levels, covering the basic authoring needs with the spatial and temporal content structures, to more comfortable authoring services with semantic content descriptions. These different authoring levels identify different abstract levels for media content descriptions. In [4], the semantics of media content are organized around *expression* and *content* levels, where the *expression* is the means used to represent the structure of media content, whereas the *content* is the representation of the conceptual aspect of the media content. Our work with the goal of providing more comfortable authoring services, proposes a model with three abstraction levels: *structure*, *concept* and *classification* (Figure 3).

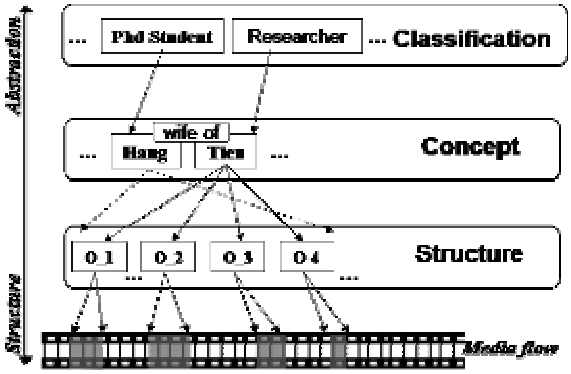

Figure 3. General model through an example of a video content description.

The example in Figure 3 shows that the structure level contains different descriptions of occurrences such as the moving regions; in the concept level these occurrences can be collected into meaning objects; finally at highest level, these meaning objects can be grouped into a known terms.

The MPEG-7 framework can support the description of all these description features. More precisely,

- The MPEG-7 audio and visual parts (part 3 and 4) can provide wide range of description tools, from generic and low level audiovisual descriptors (e.g., spectrum envelopes, fundamental frequency, GridLayout, etc.) to more sophisticated description tools like Sound effects description,

Musical instrument timbre description, Color Texture, Shape, spatio-temporal Localization, Motion, Face recognition, etc.;

- The structural aspect of the MPEG-7 Multimedia Description Schemes tools (part 5) specifies tools for describing the structure of multimedia content in time and space;
- The conceptual aspect provides the description schemes that deals with the content narrative description;
- The model description schemes tools describe parameterized models of content;
- The classification tools describe information about known collections of content in terms of labels; and so on.

Theoretically, our three abstract level models for multimedia authoring can be expressed by using the MPEG-7 description schemes. However, the expression and the experimentation for all of three levels require very complex works.

As a first step, the work presented here will focus on the first basic level, the structure description, by using the MPEG-7 structure description of media content. This level provides the spatial and temporal description data for basic authoring. A more advanced model will later interface with more semantic descriptions of MPEG-7 such as the conceptual aspect or the organization description of content.

5.2 MPEG-7 Structure Description

The MPEG-7 structural aspect supplies the generic description scheme *Segment DS* and its extensions (Figure 4 adopted from [1]) which can be used for describing the structure of multimedia content in time and space.

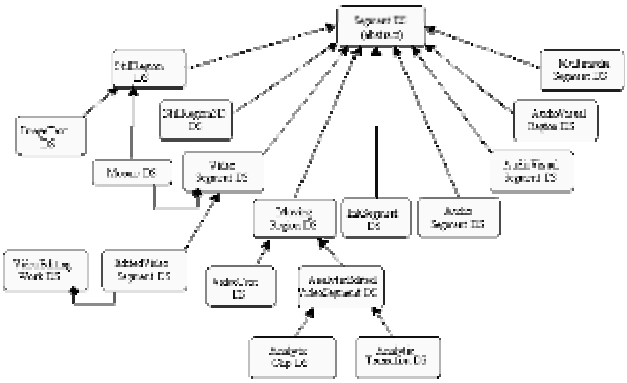

Figure 4. Hierarchical schema of the set of MPEG-7 segment description schemes.

These segment description schemes supply a rich basis for describing the structure of multimedia contents by following three global axes: *attributes*, *structural decomposition* and *relations*

- The attributes of a segment give spatio-temporal information, media information (location, creation, use, etc.), characteristics and masks, weight of importance, weight of importance given by a point of view, etc.
- The structural decomposition describes the hierarchical structure of multimedia segment.
- The relations describe structural relations among segments as temporal, spatial and spatiotemporal relations, and other conceptual relations.

The most abstract segment description schemes *Segment DS* provides almost all the common description schemes that can be applied to more specific segment schemes such as *VideoSegment DS*, *AudioSegment DS*, *StillRegion DS*, and so on. These common properties can be grouped into three parts: *indexation*, *management* and *relation*.

The *indexation* schemes of *Segment DS* supply mainly means for indexing segments. For example, the *StructuralUnit* element is used to describe the role of the segment in the structure of the multimedia contents (*information*, *summary* and *report* in a television/radio report; or *sequence*, *scene* and *plan* in a film); the *MatchingHint* element allows to associate a weight of importance to a criterion of comparison; the *TextAnnotation* element is used to put comments on the segment; finally, the *PointOfView* element defines its importance according to a specific point of view; etc. This information is very useful for search engines and even for advanced authoring tools where an integrated search engine can help in finding an appropriate media segment. However, these features are still far from our basic authoring needs. The *management* schemes of *Segment DS* such as *DescriptionMetadata* (inherited from *mpeg7:DSType*), *MediaInformation*, *MediaLocator*, *CreationInformation* and *UsageInformation* are not central for authoring activities. The *relation* schemes of *Segment DS* can allow the description of relations between segments of content. For example, the segment *A* is on the left of the segment *B*. It seems possible to use such relations for providing advanced editing functions as: starting a sound when the segment *A* meets the segment *B*. However that is out of this first experimental work.

So almost all information described by these common description schemes in the abstract tool *Segment DS* is not essential for our authoring application. It means that all information useful for authoring lies on specific segment description schemes.

In the next section, we show how we have defined our video segment and video object occurrence description models through the application of restrictions on the most complex specific segments that are *VideoSegment DS* and *MovingRegion DS*. The restriction of the other segments such as *AudioSegment DS* and *StillRegion DS* is only a simplest case of the restriction applied to *VideoSegment DS*.

5.3 Video Segment Modeling by Using MPEG-7 Video Segment Description Schemes

The *Videosegment DS* describes a temporal interval of a segment of video, which can correspond to a sequence of frames, a simple frame, or even a whole video. The video segment can be continuous or intermittent in the time. The *VideoSegment DS* provides the elements to describe *temporal*, *visual* and *structural decomposition* properties:

Temporal description schemes: the *MediaTime* element allows to localize a segment of the video in the time space by specifying its beginning instant and its duration (Figure 5a); the *TemporalMask* element allows to describe the temporal fragmentation in case the segment is discontinuous (Figure 5b).

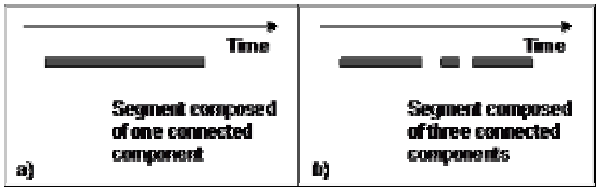

Figure 5. Examples of temporal segments

Visual description schemes: Numerous elements allow to describe the visual characteristics of a video segment. For example *VisualDescriptor*, *VisualDescriptionScheme*, *Mosaic* elements are used for describing the colour, the texture, the edge, the mosaic of video segments; the *TimeSeriesDescriptors* element is used for describing a temporal sequence with its visual characteristics.

Structural decomposition schemes supply several manners to describe the decomposition of a segment in space (*SpatialDecomposition*), time (*TemporalDecomposition*) and even both (*SpatioTemporalDecomposition*). In particular, the *MediaSourceDecomposition* element allows describing several sources in a multimedia segment. For instance, a segment of a movie can contain both video segments and audio segments.

The temporal descriptions *MediaTime* and *TemporalMask* are clearly very important for integration and synchronization authoring. In the same way, the structural decomposition descriptions are needed to describe hierarchical structure of segments. So we have selected them for our model. But even if the visual descriptions could be very useful for generic synchronizations (see the feature 3 in the section 5.1, we have not taken them in this work.

Therefore our video segment description model will refine the MPEG-7 video segment description scheme by restricting all elements inherited from the abstract segment description scheme (mpeg7:SegmentType). The XML syntax of our video segment model is as follows:

```
<element name="VideoSegment">
<complexType>
  <complexContent>
    <restriction base="mpeg7: VideoSegmentType">
      <sequence>
          <element name="MediaTime" … />
          <element name="TemporalMask" … minOccurs="0"/>
          <choice minOccurs="0" maxOccurs="unbounded">
            <element name="VisualDescriptor"… />
            <element name="VisualDescriptionScheme"…/>
            <element name="TimeSeriesDescriptors" …/>
            <element name="Mosaic" … />
          </choice>
          <choice minOccurs="0" maxOccurs="unbounded">
            <element name="SpatialDecomposition"…/>
            <element name="TemporalDecomposition"…/>
            <element name="SpatioTemporalDecomposition"…/>
            <element name="MediaSourceDecomposition"…/>
          </choice>
      </sequence>
    </restriction>
  </complexContent>
</complexType>
</element>
```

Figure 6. XML definition of our video segment model based on MPEG-7 video segment description scheme.

5.4 Occurrence Modeling by Using MPEG-7 MovingRegion Description Schemes

The MovingRegion description scheme is an extension of the *Segment DS* to describe 2D moving regions.

The *MovingRegion* description scheme also provides a rich set of description tools that can be grouped in three parts: the location (*SpatioTemporalLocator*, *SpatioTemporalMask*), the characteristic (*VisualDescriptor*, *VisualDescriptionScheme*, etc.) and the decomposition (*SpatialDecomposition*, etc.). As follows the features of our general model presented in the section 5.1, all of these descriptions of moving regions can apply to multimedia integration authoring. So we propose an occurrence model that only restricts the descriptions inherited from the abstract segment description scheme. The XML syntax of our occurrence model is refined from the MPEG-7 MovingRegion description scheme as follows:

```
<element name=="Occurence">
<complexType>
  <complexContent>
    <restriction base="mpeg7: MovingRegionType ">
      <sequence>
        <element name="SpatioTemporalLocator"
                 type="mpeg7:SpatioTemporalLocatorType" minOccurs="0"/>
        <element name="SpatioTemporalMask" …/>
        <choice minOccurs="0" maxOccurs="unbounded">
          <element name="VisualDescriptor" type="mpeg7:VisualDType"/>
          <element name="VisualDescriptionScheme" type="mpeg7:VisualDSType"/>
          <element name="TimeSeriesDescriptors" type="mpeg7:TimeSeriesType"/>
        </choice>
        <choice minOccurs="0" maxOccurs="unbounded">
          <element name="SpatialDecomposition"…./>
          <element name="TemporalDecomposition"…/>
          <element name="SpatioTemporalDecomposition"…/>
          <element name="MediaSourceDecomposition"…/>
        </choice>
      </sequence>
    </restriction>
  </complexContent>
</complexType>
</element>
```

Figure 7. XML definition of our occurrence model based on MPEG-7 MovingRegion description scheme.

However, only the *SpatioTemporalLocator* description is experimented by now, because it allows to edit the basic fragment integration such as spatio-tempotral synchronizations with occurrences (a text element follows an occurrence of a video object); hyperlinks on occurrences of a video object; or even extract occurrences of an object from the media to integration. More detail about these applications are given in section 7.

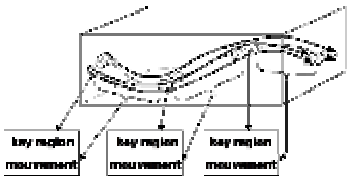

Figure 8. spatio-temporal region (adopted from [5])

In fact, The *SpatioTemporalLocator* describes spatio-temporal regions in a video sequence and provides localization functionality especially for hypermedia applications. It consists of *FigureTrajectory* and *ParameterTrajectory* that describe a set of reference regions and their motions (Figure 8 adopted from [5]).

5.5 Synthesis

A general content description model with its features for authoring has been specified in this section. Only the most basic structures of content have been selected in this first work. The model is expressed by restricting the structural description schemes of MPEG-7. By this way, our model can interoperate with MPEG-7 description data. In addition, it only interfaces with the subset of description data that is necessary for our authoring

application. That will reduce the complexity of implementing MPEG-7 description data. Next section explains how this model is used in a multimedia integration model.

6. MULTIMEDIA FRAGMENT INTEGRATION MODEL

An important issue is how to use the description data in integration model for authoring multimedia presentations. As shown in sections 3 and 4 the difficulty relies on the gap between multimedia content description and multimedia integration models. Our previous work described in [12] has provided an extension model of multimedia integration to overcome this gap. In this section, we only explain the general principles of that work.

A new media type is added in the integration model. It is the *structured media* type including *structured video*, *structured audio*, *structured text*, *structured image*, etc. The *Structured media* type is a media node in a document that not only refers to a media content resource, but also can be included or can refer to a structure description of the content. The structure description will help the authoring and presentation system to visualize the structure of the content in time and in space and then will make the fragment integration and synchronization easier (Figure 9).

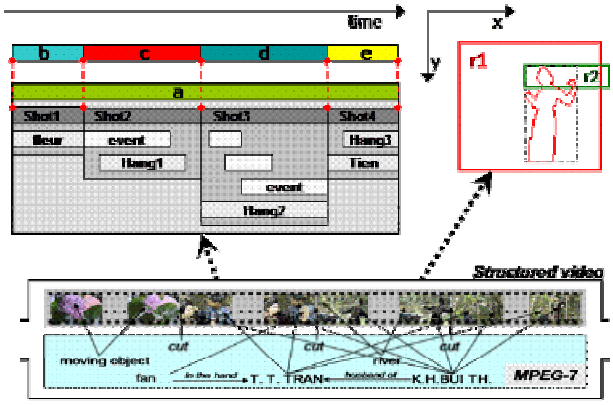

Figure 9. Multimedia fragment integration authoring with a structured video.

For instance, the author wants a sequence of images to be synchronized with a sequence of fragments (shots) from a structured video. The temporal presentations *b*, *c*, *d* and *e* of the images must be synchronized with the boundaries of cut shots of the video. Similarly, the spatial presentation *r2* of an object needs to be always on the top of a moving region in the video. Note that the temporal and spatial presentations of the video are represented through the bar *a* and the rectangle *r1*. The authoring of such a multimedia integration will be much easier if the video used in the document is a *Structured Video* element (Figure 9).

Next section describes how these descriptions will be experimented in an authoring environment of multimedia document.

7. AUTHORING APPLICATION

The experimentation is performed in a multimedia integration authoring and presentation tool called *Mdefi* [12].

Mdefi adopted the WYSIWYG approach for editing and presenting multimedia documents. It provides a multi-view environment in which the most important views are the *TimeLine*, the *Execution* and the *Hierarchical* views. The *TimeLine* and *Execution* views (Figure 10a and Figure 10b) provide the graphical interface for editing the temporal and spatial structures of document, while the *hierarchical* view (Figure 10c) offers a tree view of all structures of the document. Figure 10 presents an instance of these synchronized views in which the presentation of four images are showed.

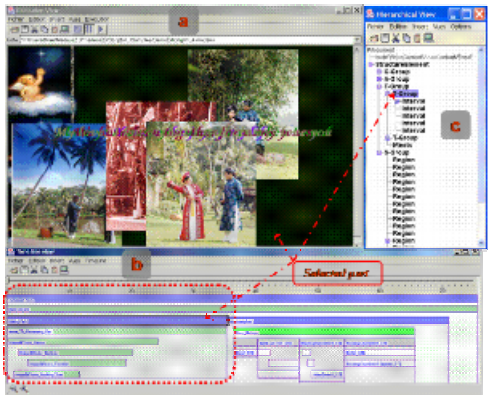

Figure 10. Multi-views interface of *Mdefi*.

Mdefi proposes also a set of views to edit structured media before their integration in a document. These new views themselves form also a multi-view system, which makes the visualization, the navigation and the manipulation on the structures of media content easier. Figure 11 presents the interface of the structured video view, in which the manipulation on a moving region is visualized in different views of video content: the hierarchical structure (1), the attribute (2) the temporal structure (3).

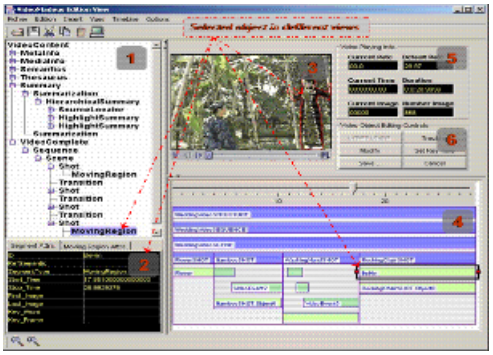

Figure 11. Structure video view.

The most important thing is that these structured media views are synchronized with the other views of the authoring system. So any segment of the media structure can be chosen to be integrated into the document being edited. The fragment integration can be done simply by a drag and drop of the described segment from a structured media view into the document *Execution* view.

If the integrated segment has itself temporal and spatial description structures, these structures will be shown in the *TimeLine* and the *spatial* views of the document. The fine-grained synchronizations with the integrated segment can be done easily thanks to this visualization. The Figure 12 shows such a case. A

scene segment of the video has been chosen to be integrated into the document. Its temporal structure is showed in the *TimeLine* view (Figure 12b), so the fine-grained synchronizations of its four shots with four images (Figure 10) can be done in a straightforward manner.

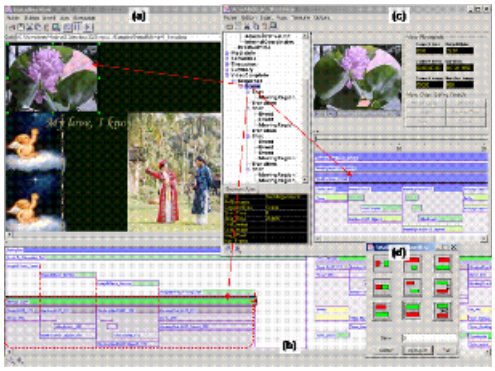

Figure 12. Temporal fine-grained synchronization authoring.

Another interesting situation occurs when the integrated segment is a moving region. The moving region will be extracted in real-time from the video and directly integrated in the edited document. Figure 13 shows such a case where a video object is inserted on the right top of the image by just a simple drag and drop action.

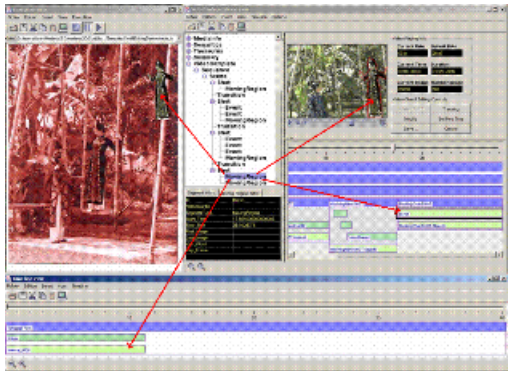

Figure 13. Video object integration.

8. CONCLUSION

This work has provided a new way for increasing the capacities of media fragment integration in the multimedia integration and synchronization authoring. It proposed the use of multimedia content description for visualization and manipulation with the media content structure. The description model is adopted from the content structure description schemes of the standard MPEG-7. That makes the system interoperable with existing MPEG-7 description data. In addition, the restriction models following the specified features can help the system to reduce the complexity of MPEG-7 description data. Finally, the experimental editing application has showed the interesting power of this approach.

However, it is only the first step of work in which we have only experimented with the structure of media content. Further experimental works with the concept, the model and the classification even with the audiovisual characteristics of media content can promise much more interesting and powerful results.

9. REFERENCES

[1] P. v. Beek, A. B. Benitez, J. Heuer, J. Martinez, P. Salembier, Y. Shibata, J. R. Smith, T. Walker, Text of 15938-5 FCD Information Technology – Multimedia Content Description Interface – Part 5 Multimedia Description Schemes, March 2001, Singapore.

[2] J.Carrive et al., "Using Description Logics for indexing Audiovisual Documents", Int. Workshop on description Logics, pp. 116-120, 1998.

[3] A. Celentano, O. Gaggi, “A Synchronization Model for Hypermedia Documents Navigation”, ACM Symposium on Applied Computing 2000

[4] Chatman, S. Story and Discourse: Narrative Structure in Fiction and Film. New York: Ithaca (1978).

[5] L. Cieplinski, M. Kim, J. Ohm, M. Pickering, A. Yamada, Text of ISO/IEC 15938-3/FCD Information Technology – Multimedia Content Description Interface – Part 3 Visual, Singapore, March 2001.

[6] J. Hunter, S. Little. *Building and Indexing a Distributed Multimedia Presentation Archive using SMIL*. ECDL '01, Darmstadt, September 2001.

[7] J. Hunter, L. Armstrong, "A Comparison of Schemas for Video Metadata Representation", WWW8, Toronto, May 10-14, 1999.

[8] L. H. Hsu et al.,”A Multimedia Authoring-in-the-Large Environment to Support Complex Product Documentation”, Multi-media Tools and Application 8, 11-64 (1999).

[9] C.-S. Li et al., “Multimedia Content Description in the InfoPyramid”, IEEE Inter. Conf. on Acoustics, Speech and Signal Processing (ICASSP-98), June, 1998.

[10] L. Rutledge, P. Schmitz, “Improving Media Fragment Integration in Emerging Web Formats”, the 8th International Conference on Multimedia Modeling (MMM 2001), CWI, Amsterdam, The Netherlands, November 5-7, 2001.

[11] J. Saarela, “Video Content Models based on RDF”, W3C workshop on “Television and the Web”, Sophia-Antipolis, France, June 1998.

[12] T. Tran_Thuong, C. Roisin. “Media content modelling for Authoring and Presenting Multimedia Document”. A book chapter in Web Document Analysis: Challenges and Opportunities, World Scientific, 2002.

[13] T. Wahl and K. Rothermel, “Representing Time in Multimedia-Systems”, Proceedings of IEEE Conference on Multimedia Computing and Systems, May 1994.

[14] Z. Zelenika, “CARNet Media on Demand - Metadata model”, web edition 2001-05-21.